Applied Causal Inference Powered by ML and AI

Victor Chernozhukov* Christian Hansen† Nathan Kallus‡
Martin Spindler§ Vasilis Syrgkanis¶

July 28, 2024

Publisher: Online
Version 0.1.1

* MIT
† Chicago Booth
‡ Cornell University
§ Hamburg University
¶ Stanford University

5 Causal Inference via Conditional Ignorability

"compare apples and/to/with apples: to compare things that are very similar."

– Merriam Webster Dictionary [1].

Here we discuss how average causal effects may be identified using regression when treatment is not randomly assigned but instead depends on observed covariates. We discuss the conditional or adjustment method, which relies on comparing the average difference between expected outcomes for treated and untreated units that are comparable (formally, identical) in terms of their characteristics X. If treatment is as good as randomly assigned conditional on X, then this approach recovers average causal or treatment effects. This key condition is commonly referred to as conditional ignorability, conditional exogeneity, or unconfoundedness.

5.1 Introduction

In a cross-country analysis, higher chocolate consumption predicts a higher number of Nobel laureates per capita.

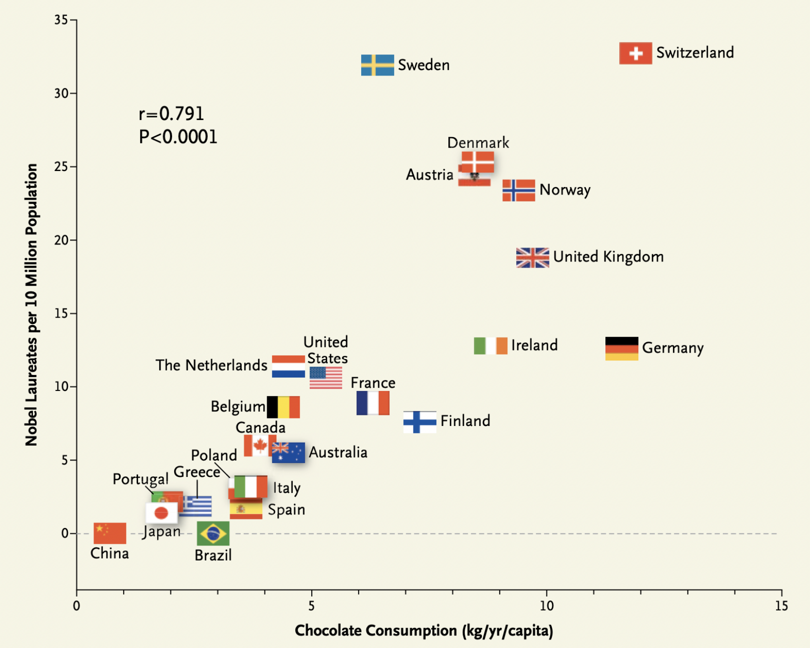

Figure 5.1: Source: Franz H. Messerli, "Chocolate Consumption, Cognitive Function, and Nobel Laureates," New England Journal of Medicine. 2012

Is this a reflection of a true causal effect and therefore an actionable insight? If it were, countries could generate more Nobel laureates per capita by making chocolate abundant to everyone. (This wouldn't be a bad thing.) Is this perhaps what Switzerland did? Switzerland has the highest number of Nobel laureates per capita.

Or is there a common cause[1] that creates non-causal association? Perhaps wealthy countries invest more in science and higher wealth causes people to consume luxury goods like chocolate. See for instance plots (D) and (E) in Figure 5.3. Comparative analysis, where we compare nations with identical or similar wealth, would probably reveal that the correlation is not causal.[2] Probably we should be comparing Switzerland to similar countries in terms of wealth – the "apples-to-apples" comparison, so to speak. This type of analysis is very common in causal

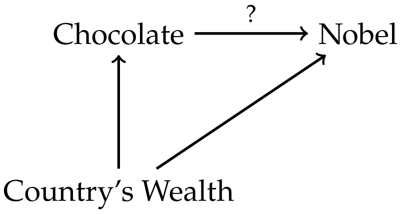

Figure 5.2: A Contrived Causal Path Diagram for the Effect of Country's Wealth on Chocolate Consumption and Nobel Prize Production per capita.

[1] We often refer to these common causes as "omitted variables" that give rise to "omitted variable bias."

[2] It remains a fundamental empirical problem to confirm this conjecture or disprove this conjecture. The causal channel through which chocolate (and other flavonoids) may affect Nobel production is by documented improvement in the cognitive function.

inference and is implemented via a set of tools introduced in this chapter.

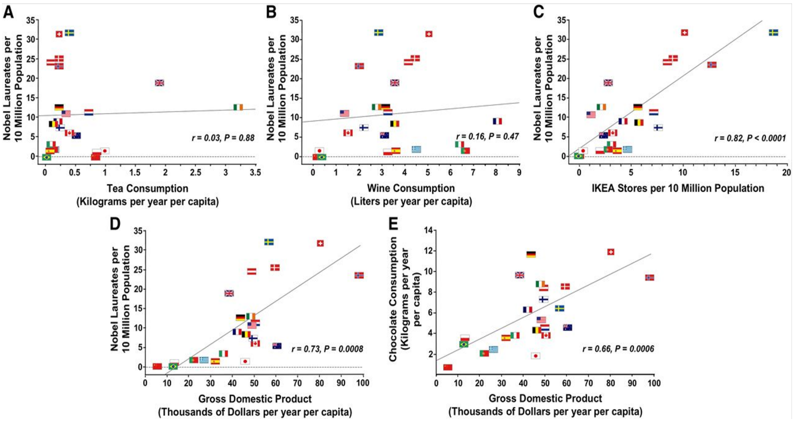

Figure 5.3: Source: J Nutr, Volume 143, Issue 6, June 2013, Pages 931–933, "Does Chocolate Consumption Really Boost Nobel Award Chances? The Peril of Over-Interpreting Correlations in Health Studies," ©2013 American Society for Nutrition

In what follows, we work within Rubin's [2] potential outcomes framework, as introduced in Chapter 2. The idea is that if we can think of observed treatment D as generated randomly – independently of potential outcomes – conditional on some pre-treatment variables X, then we can learn the average causal (treatment) effects by regression

$$\text{of } Y \text{ on } D \text{ and } X,$$

or, as is often said, by "adjusting" or "controlling" for X.

Notation

Recall that we denote the independence of two random variables (these can include random vectors) U and V as

$$U \perp\!\!\!\perp V.$$

Independence, conditional on a third variable X, is denoted by

$$U \perp\!\!\!\perp V \mid X.$$

5.2 Potential Outcomes and Ignorability

Recall that we use $Y(d)$ to denote potential outcome in the treatment state d, where we consider only the case $d \in \{0, 1\}$ for simplicity. We also recall our example of smoking from Chapter 2. Suppose we want to study the impact of smoking marijuana on life longevity. Suppose that smoking marijuana has no causal/treatment effect on life longevity:

$$Y = Y(0) = Y(1), \text{ so that } \delta = \mathrm{E}[Y(1)] - \mathrm{E}[Y(0)] = 0.$$

However, the observed smoking behavior, D, results not from an experimental study, but from observational data in which an individual's smoking decisions are driven by other behavioral choices X (drinking alcohol for example) which cause shorter life longevity. In this case, the predictive effect recovered by regression without adjusting for X does not match the average causal effect

$$\mathrm{E}[Y \mid D = 1] - \mathrm{E}[Y \mid D = 0] < 0 = \delta,$$

because higher D predicts higher X, which predicts lower Y. This difference between the predictive effect and average causal effect is the result of confounding or *selection bias*.

In this example, conditioning on X can remove the selection bias (see Figure 5.4)

$$\mathrm{E}[\mathrm{E}[Y \mid D = 1, X] - \mathrm{E}[Y \mid D = 0, X]] = \delta,$$

provided that conditional on X variation in D is independent of the potential health outcomes.

The following provides a formal assumption under which we can eliminate the confounding bias by controlling for X.[3]

Assumption 5.2.1 (Conditional Ignorability and Consistency) *Ignorability: Suppose that treatment status D is independent of potential outcomes $Y(d)$ conditional on a set of covariates X: For each d,*

$$D \perp\!\!\!\perp Y(d) \mid X.$$

Consistency: Suppose that Y is generated as $Y := Y(D)$.

Identification by Conditioning

The ignorability assumption[4] says that variation in treatment assignment D is as good as random conditional on X. This assumption means that if we look at units with the same value of the covariates, e.g. units with $X = x$, then treatment variation among these observationally identical units, $D \mid X = x$, is indeed produced as if by a formal randomized control trial.

[3] 3: The assumption is fundamentally untestable and is an assumption in the purest sense. Given assumed domain knowledge encoded in causal DAGs, we study a systematic way of finding X that satisfy this assumption in subsequent chapters.

[4] 4: You may wonder why the term "ignorability" is used. The distribution of $Y(d)$ depends only on X and not on D, so the latter is "ignorable." Note that the conventional name used in econometrics for the ignorability assumption is the *conditional exogeneity* or *conditional independence* assumption.

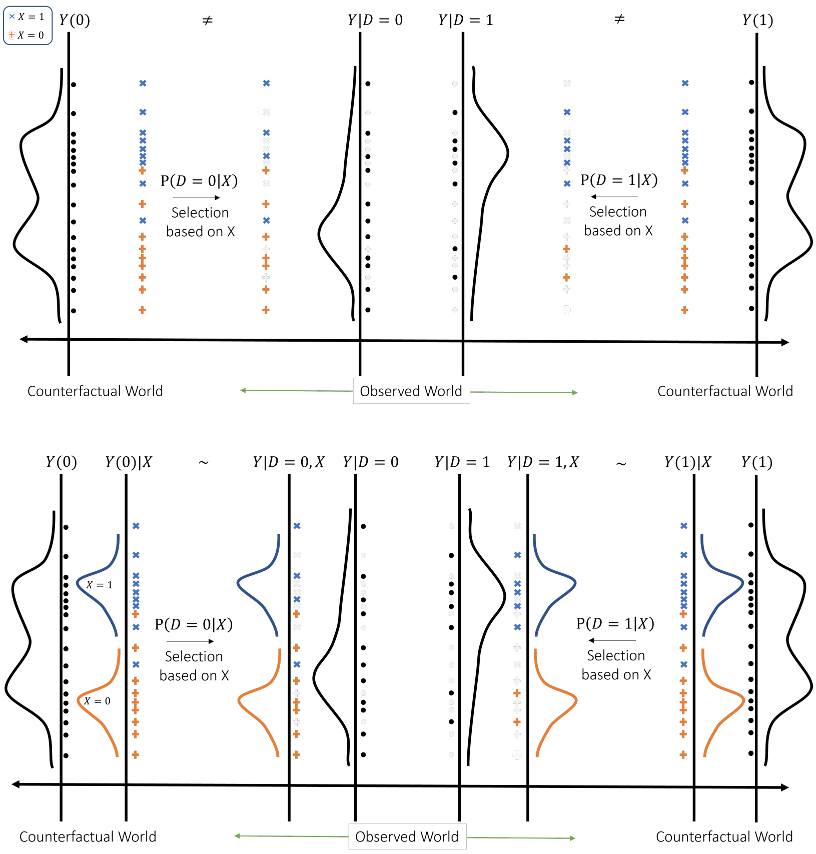

Figure 5.4: Pictorial representation of how selection on X can lead to biased observed outcomes between treated and control populations, while conditioning on X removes the selection bias. In this example, the potential outcomes $Y(0)$ and $Y(1)$ have identical distributions shown in the far left and right of the figure. We also have a binary covariate X that is related to treatment probability in the sense that $\mathrm{P}(D = 1|X = 1) > \mathrm{P}(D = 1|X = 0)$ and $\mathrm{P}(D = 0|X = 1) < \mathrm{P}(D = 0|X = 0)$ which leads to selection bias when we do not condition on X. This bias is illustrated by the difference in the distribution of (observed) Y given $D = 0$ and $D = 1$ shown in the black curves in the middle of the figure. The bottom panel then shows that selection bias is removed by conditioning on X as the distribution of potential outcomes given X (blue and orange curves under $Y(0)|X$ and $Y(1)|X$) equals the distribution of observed outcomes given D and X (blue and orange curves under $Y|D = 0, X$ and $Y|D = 1, X$).

Therefore, we can learn about the causal effect of D by comparing outcomes across treated and control units who have identical characteristics $X = x$ under the conditional ignorability assumption. The idea of comparing observations who have identical characteristics is the essence of the so-called *conditioning* or *adjustment* strategy to learning causal effects. As conditioning approaches produce a different contrast for every potential value of X, we may also wish to average the contrasts at different values of X over the distribution of characteristics to produce a summary measure of the causal effects.

The conditional probability of receiving treatment, *the propensity score*, plays an important role in this approach.

Assumption 5.2.2 (Overlap/Full Support) *The probability of receiving treatment given X, the propensity score*

$$p(X) := \mathrm{P}(D = 1|X),$$

is non-degenerate:

$$\mathrm{P}(0 < p(X) < 1) = 1.$$

The overlap assumption requires that there is proper randomization or variation in D at each value x in the support of X. Without this condition, there are values x in the support of X where we cannot construct a contrast between treatment and control units. We cannot learn the conditional average treatment effect at these values of X and thus are also unable to learn the unconditional average effect of the treatment.

Remark 5.2.1 Assumption 5.2.2 is also often called the *full support* condition because it requires

$$\mathrm{support}(\mathrm{D}, \mathrm{X}) = \{0, 1\} \times \mathrm{support}(X).$$

The following is the most important theoretical result that states that we can recover expectations of potential outcomes from regressions.

Theorem 5.2.1 (Conditioning on X Removes Selection Bias) *Under Conditional Ignorability and Overlap, the conditional expectation function of observed outcome Y given $D = d$ and X recovers the conditional expectation of the potential outcome $Y(d)$ given X:*

$$\mathrm{E}[Y \mid D = d, X] = \mathrm{E}[Y(d) \mid D = d, X] = \mathrm{E}[Y(d) \mid X].$$

To prove Theorem 5.2.1, note that the overlap assumption makes it possible to condition on the events $\{D = 0, X\}$ and $\{D = 1, X\}$ at any value in the support of X and that the second equality holds by ignorability.

Hence, the Conditional Average Predictive Effect (CAPE),

$$\pi(X) = \mathrm{E}[Y \mid D = 1, X] - \mathrm{E}[Y \mid D = 0, X],$$

is equal to the Conditional Average Treatment Effect (CATE),

$$\delta(X) = \mathrm{E}[Y(1) \mid X] - \mathrm{E}[Y(0) \mid X].$$

Thus, the APE and ATE also agree:

$$\delta = \mathrm{E}[\delta(X)] = \mathrm{E}[\pi(X)] = \pi.$$

Conditional Ignorability via Causal Diagrams

It is possible to illustrate the key ignorability assumption, Assumption 5.2.1, graphically as follows:[5]

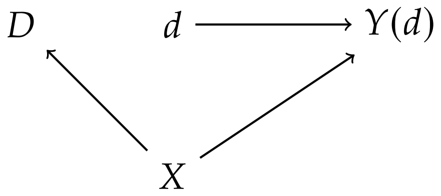

Figure 5.5: A Causal Diagram for the Conditional Ignorability Research Design

In this graph, we show the potential outcome $Y(d)$ as a node and the potential treatment status d as another node. The latter node is deterministic. There is an arrow from d to $Y(d)$ indicating the dependency. The pre-treatment covariates X affect both the realized treatment variable D and the potential outcomes $Y(d)$, as shown by the arrow from X to D and from X to $Y(d)$. The assigned treatment variable D is independent of the node $Y(d)$, conditional on X. Independence can be derived from the graph by observing the absence of any path between the D and $Y(d)$ nodes other than the path through the variable X upon which we've conditioned. Note that Assumption 5.2.2, the overlap condition, is not illustrated in the graph.

The potential outcome process $d \mapsto Y(d)$ and treatment assignment jointly determine the realized outcome variable Y via the assignment $Y := Y(D)$. This generates the following causal diagram. This graph says that X is generated first. D is then generated, with the distribution of D depending on X. Finally, Y is generated, with its distribution depending on both D and X. Here, after conditioning on X, the statistical dependence (association) between D and Y only reflects the causal channel, $D \to Y$ allowing us to uncover the ATE, for example.

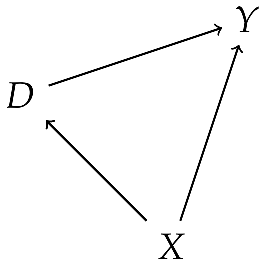

Figure 5.6: A Causal Diagram with Conditional Ignorability

[5] Note that what we present is just one of many casual diagrams that are compatible with the conditional ignorability condition. There are others, as will become apparent in subsequent chapters.

Connections to Linear Regression

The tools from Chapter 1 and Chapter 4 can be used to perform statistical inference on ATEs. We briefly discuss how (high-dimensional) regression can be used to retrieve causal estimates when conditional ignorability holds in this section.

The simplest instance of the problem is when the conditional expectation function of Y given D and X is linear,

$$\mathrm{E}[Y \mid D, X] = \alpha D + \beta' W,$$

which gives a model

$$Y = \alpha D + \beta' W + \epsilon, \quad \mathrm{E}[\epsilon \mid D, X] = 0.$$

Here it is understood that W may include X as well as pre-specified nonlinear transformations of X.

In this model, α identifies δ

$$\delta = \alpha$$

under the linearity assumption and ignorability, and our inference tools for α automatically carry over to δ. Note that the linearity assumption and ignorability assumptions imply that treatment effects are homogeneous; that is, $\delta(x) = \delta$ for all x in the support of X.

Of course, the assumption of linearity and homogeneous treatment effects is restrictive. A simple way to relax this is to consider interactions. One version of this approach takes all interactions between W and D and assumes

$$\mathrm{E}[Y \mid D, X] = \alpha_1 D + \alpha_2' W D + \beta_1 + \beta_2' W,$$

where we also maintain that we are working with centered covariates: $\mathrm{E}W = 0$.[6]

We then recover the ATE as

$$\delta = \alpha_1$$

and CATE as

$$\delta(X) = \alpha_1 + \alpha_2' W.$$

We can use partialling out methods, such as OLS in the low-dimensional case and Double Lasso (and variants) in

[6] This model is still linear and results for linear models carry over to this case as well.

the high-dimensional case, to perform inference on α_1 and components of α_2. We can use these same methods to perform inference over β_1 and components of β_2, though these parameters will often not be of interest.

Note, we used this approach in the heterogeneous wage gap example in Chapter 1. The discussion of whether the wage gap analysis has a causal interpretation is given in the next causal inference chapter, Chapter 6.

As demonstrated in Theorem 5.2.1, the ultimate targets are the conditional expectation functions $\mathrm{E}[Y(d)|X]$ if our goal is to learn average causal effects under ignorability. This being our target makes the relevance of considering transformations $W = T(X)$ of X important as we would like to have the linear model provide a good approximation to these conditional expectation functions. See the discussion in "From Best Linear Predictor to Best Predictor" in Chapter 1. If the linear model is misspecified in the sense that it does not approximate the conditional expectation functions well, the estimated causal effects - e.g. α_1 in the interactive model - do not necessarily have any causal interpretation. This potential failure is a major reason we consider more flexible, modern machine learning methods.

What about fully nonlinear strategies? We will explore them in Chapter 10.

5.3 Identification Using Propensity Scores

The identification by conditioning approach requires being able to accurately model the "outcome process," i.e. the conditional expectation function $\mathrm{E}[Y \mid D, X]$. This conditional expectation function might correspond to a complicated real world process that is hard to model or approximate.

When the outcome process is hard to model, we might have a much better handle on the "treatment selection process," i.e. the propensity score:

$$p(X) = \mathrm{P}(D = 1 \mid X).$$

An alternative approach, known as the Horvitz-Thompson method [3], uses propensity score reweighting to recover averages of potential outcomes. Using the propensity score rather than identification by conditioning on X is a useful empirical strategy when X is high-dimensional and $p(X)$ is available or can be approximated accurately.[7] An example of a setting where the propensity score is known is a *stratified RCT*, which is an experiment where treatment is assigned at random with probability $p(X)$ to individuals with different observed covariates X. In this case, the treatment assignment probability $p(X)$ is exactly the propensity score.

Theorem 5.3.1 (Horvitz-Thompson: Propensity Score Reweighting Removes Bias) *Under Conditional Ignorability and Overlap, the conditional expectation of an appropriately reweighted observed outcome Y, given X, identifies the conditional average of potential outcome $Y(d)$ given X:*

$$\mathrm{E}\left[Y \frac{1(D=d)}{\mathrm{P}(D=d|\mathrm{X})} \mid X\right] = \mathrm{E}[Y(d) \mid X]$$

Then, averaging over X identifies the average potential outcome:

$$\mathrm{E}\left[Y \frac{1(D=d)}{\mathrm{P}(D=d|\mathrm{X})}\right] = \mathrm{E}[Y(d)]$$

To prove this result, note

$$\begin{aligned}\mathrm{E}\left[Y \frac{1(D=d)}{\mathrm{P}(D=d|\mathrm{X})} \mid X\right] &= \frac{\mathrm{E}[Y1(D=d) \mid X]}{\mathrm{P}(D=d|\mathrm{X})} \\ &= \mathrm{E}[Y(d) \mid X]\frac{\mathrm{E}[1(D=d) \mid X]}{\mathrm{P}(D=d|\mathrm{X})} \\ &= \mathrm{E}[Y(d) \mid X],\end{aligned}$$

where we used conditional ignorability in the second equality.

As a consequence, we can identify average treatment effects by simple averaging of transformed outcomes:

$$\delta = \mathrm{E}[YH], \quad H = \frac{1(D=1)}{\mathrm{P}(D=1|X)} - \frac{1(D=0)}{\mathrm{P}(D=0|X)},$$

where H is called the Horvitz-Thompson transform. Similarly, we can identify conditional average treatment effects as a conditional average of transformed outcomes:

$$\delta(X) = \mathrm{E}[YH \mid X].$$

[7] An interesting example where the propensity score is not known but can be well-approximated is the examination in [4] of the causal effect of attendance at a particular school or group of schools relative to one or more alternative schools (e.g., "elite" vs. "non-elite" schools) in settings where matching algorithms are used to assign students to schools. In this example, we can think of these student assignment mechanisms as $p(X)$.

Note that propensity score reweighting reduces to the difference of means in the control and treatment groups when the propensity score is constant.

Stratified RCTs

In the case where the propensity score $p(X)$ is known, we are essentially back to a classical RCT.

Definition 5.3.1 (Generalized/Stratified RCT) *If under Assumption 5.2.1, the propensity score $p(X)$ is known, the setting is called a generalized or stratified RCT.*

Remark 5.3.1 Propensity score reweighting is generally not the most efficient approach to estimating treatment effects from a statistical point of view because it ignores any dependence between the outcomes and controls, X, that is not captured by the propensity score. By exploiting dependence between the outcomes and X not captured by the propensity score, more efficient estimation of treatment can occur as using this dependence "de-noises" the outcome. Moreover, estimation based on only propensity score reweighting fails under imbalances that might arise due to imperfect data collection. Later, we will use *both* regression and reweighting as part of "double machine learning" to operationalize efficient statistical inference on treatment effects in fully nonlinear (nonparametric) models.

Covariate Balance Checks

Given a propensity score $p(X)$, we can check if the RCT is valid (randomization is successful) by performing a *covariate balance check.*. Specifically, conditional ignorability implies that

$$\mathrm{E}[H \mid X] = 0.$$

Thus, if covariates predict H, we can conclude that conditional ignorability does not hold. Heuristically, covariates predicting H means that covariates are imbalanced in the sense that, after reweighting by X dependent treatment probability, there are systematic differences in X across treatment and control observations which can be exploited to predict treatment assignment.

In a low-dimensional linear model framework, a covariate balance check can be done by regressing H on W, a dictionary of transformations of X, and testing if W predicts H. W predicting H suggests that the RCTs randomization protocol did not go as planned.

Connections to Linear Regression

Note that by the Horvitz-Thompson transform characterization of the CATE, $\delta(X) = \mathrm{E}[YH \mid X]$, we can view the conditional average treatment effect as the solution to a prediction problem of predicting the transformed outcome YH from the regressors X.

A useful strategy is to consider (potentially high-dimensional) linear regression models where HY is the dependent variable; see, e.g., [5]. Note that if we assume that $\mathrm{E}[Y \mid D, X] = \alpha_1 D + \alpha_2' WD + \beta_1 + \beta_2' W$, where W is a dictionary of transformations of X, then we have

$$\mathrm{E}[YH \mid X] = \alpha_1 + \alpha_2' W.$$

Thus, we can simply run a regression of YH on $(1, W')'$. In this regression model, we recover the ATE as

$$\delta = \alpha_1$$

and CATE as

$$\delta(X) = \alpha_1 + \alpha_2' W.$$

We can use partialling out methods, such as Double Lasso, to perform inference on α_1 and components of α_2. We also discuss estimating CATE using more general machine learning methods in Chapter 14 and Chapter 15.

5.4 Conditioning on Propensity Scores*

The fact that conditioning on the right set of controls removes selection bias has long been recognized by researchers employing regression methods. Rosenbaum and Rubin [6] made the much more subtle point that conditioning on only the propensity score

$$p(X) = \mathrm{P}(D = 1 \mid X)$$

also suffices to remove the selection bias.

Theorem 5.4.1 (Rosenbaum and Rubin: Conditioning on the Propensity Score Removes Selection Bias) *Under Ignorability and Overlap, D is generated independently of $Y(d)$ for each d, conditional on the propensity score $p(X)$: For each d,*

$$D \perp\!\!\!\perp Y(d) \mid p(X).$$

In other words, conditional on $p(X) = p$, variation in D is as good as randomly assigned. Hence, whenever it suffices to use X for identification by conditioning, it also suffices to use $p(X)$. This fact makes $p(X)$ a "minimal sufficient" statistic, conditioning on which removes selection bias under ignorability.

In scenarios with a known propensity score, we can simply use $p(X)$ as a control in place of the high-dimensional set of characteristics, X, and thus bypass a potentially complicated high-dimensional estimation problem. In other words, we can identify the conditional average potential outcome as

$$\mathrm{E}[Y(d) \mid p(X)] = \mathrm{E}[Y \mid D = d, p(X)].$$

Thus, it suffices to learn the CEF $\mathrm{E}[Y \mid D, p(X)]$. We learn good approximations of these CEFs by incorporating polynomials or other transformations of $p(X)$ to make things more flexible and running linear regression methods. Finally, we can also employ nonlinear machine learning methods introduced in Chapter 9 to overcome the limitations of linear models.

After controlling for $p(X)$, we can also consider the use of high-dimensional methods to include other transformations W of the raw variables X in order to improve precision, estimating the more flexible CEF $\mathrm{E}[Y \mid D, p(X), W]$. It is especially advisable to include transformations W that fail the covariate balance checks discussed in Section 5.3. Including W can reduce the selection bias (and, hopefully, set it equal to zero). In the reemployment experiment, for example, we observed that balance did not seem satisfied across age groups. Hence, further controlling for age makes sense and results in modest changes to estimates of the treatment effect. Of course, there is no guarantee that controlling for observed covariates can overcome selection bias in compromised RCTs in general because unobserved covariates may be driving the bias.

Remark 5.4.1 ("Clever Covariate") Finally, we note that the simple OLS regression of Y on the single constructed regressor

$$\phi(D, X) := \frac{1(D = 1)}{\mathrm{P}(D = 1|X)} - \frac{1(D = 0)}{\mathrm{P}(D = 0|X)} = H$$

can be used to estimate the ATE. Specifically, for β the coefficient in the model $Y = \beta H + \varepsilon$ with $\varepsilon \perp H$, we have that the ATE is equal to $\mathrm{E}[\beta(\phi(1, X) - \phi(0, X))]$. This result holds even though the CEF function is not given by βH; see Section 5.B. As such, incorporating the technical regressor H in a linear regression model (without penalization if high-dimensional estimation tools are used) can be a good idea. This approach is referred to as the "clever covariate" approach in the literature [7, 8].

5.5 Average Treatment Effect for Groups and on the Treated

In addition to unconditional average treatment effects (ATE) or average treatment effects at specific values of the covariates $X = x$, we may be interested in average effects within specific subpopulations.

A leading example of an interesting subpopulation treatment effect is a group ATE (GATE):

$$\delta_G = \mathrm{E}[Y(1) - Y(0)|G = 1]$$

where G is a group indicator defined in terms of X's. For example, we might be interested in the effects of a training program among younger people, say between 18 and 30 years old ($G = 1(18 \leq \texttt{age} \leq 30)$); among people older than 30 years old (so $G = 1(30 < \texttt{age})$); and differences between these two groups.

We can immediately obtain the GATE using the identification results above and the law of iterated expectations:

$$\begin{aligned}
&\mathrm{E}[Y(1) - Y(0)|G = 1] \\
&\qquad = \mathrm{E}[\mathrm{E}[Y|D = 1, X] - \mathrm{E}[Y|D = 0, X]|G = 1] \\
&\qquad = \mathrm{E}[HY|G = 1].
\end{aligned}$$

That is, we can identify GATEs either by taking the difference in regression functions or applying propensity score reweighting of outcomes and then averaging over group G.

We next consider treatment effects for the subpopulation of treated units, the *average treatment effect on the treated* (ATET):[8]

$$\delta_1 = \mathrm{E}[Y(1) - Y(0) \mid D = 1].$$

For example, consider training completion as a treatment, D, and X a vector of pre-treatment variables such that unconfoundedness holds. Consider the question:

- On average, how much more do trainees earn after going through the training program than they would have earned had they not gone through the program?

Note that this question is a counterfactual question as it requires us to compare outcomes for trainees in the treated state, where they receive training, and the unobserved control state, where they did not receive training. The ATET, δ_1, is the parameter that answers such questions about counterfactuals. The ATET is identified by

$$\mathrm{E}[\mathrm{E}[Y|D = 1, X] - \mathrm{E}[Y|D = 0, X] \mid D = 1]$$

similarly to what we had above. It is also possible to bypass the use of $\mathrm{E}[Y|D = 1, X]$ in this case; see Appendix 5.C for more details.

Study Problems

1. Use one or two paragraphs to explain conditioning and its use in learning treatment effects/causal effects in observational data and randomized trials where treatment probability depends on pre-treatment variables. This discussion should be non-technical as if you were writing an explanation for a smart friend with relatively little exposure to causal modeling.

2. Use one or two paragraphs to explain the propensity score reweighting approach for identification of average treatment effects. This discussion should be non-technical as if you were writing an explanation for a smart friend with relatively little exposure to causal modeling.

3. Use one or two paragraphs to explain why group ATE and the ATE on the treated may be of interest in empirical work. This discussion should be non-technical as if you were writing an explanation for a smart friend with

[8] Rather than ATET, some use the abbreviation AToT or ATT.

relatively little exposure to causal modeling.

5.A Rosenbaum-Rubin's Result

Recall the propensity score is

$$p(X) := \mathrm{P}(D = 1|\mathrm{X}),$$

which is the probability of receiving treatment given X. A simple useful intermediate property is the balancing property of the propensity score which states that treatment is independent of X conditional on the propensity score:

$$D \perp\!\!\!\perp X \mid p(X) \quad \Leftrightarrow \quad \mathrm{P}(D = 1|X, p(X)) = \mathrm{P}(D = 1|p(X)).$$

This result follows simply from (i) $\mathrm{P}(D = 1|X, p(X)) = \mathrm{P}(D = 1|X) = p(X)$ and (ii) $\mathrm{P}(D = 1|p(X)) = \mathrm{E}[D = 1|p(X)] = \mathrm{E}[\mathrm{E}[D|X, p(X)]|p(X)] = \mathrm{E}[p(X)|p(X)] = p(X)$. This property underlies covariate balance checks.

We now turn to the theorem of Rosenbaum and Rubin. By Theorem 5.3.1 and the law of iterated expectations, we have that for any function of the form $g(y) = 1(y \leq t), t \in \mathbb{R}$:

$$\begin{aligned}
\mathrm{E}\left[g(Y(1)) \mid p(X)\right] &= \mathrm{E}[\mathrm{E}[g(Y(1))|X, p(X)]|p(X)] \\
&= \mathrm{E}[\mathrm{E}[g(Y(1))|X]|p(X)] \\
&= \mathrm{E}\left[g(Y)\frac{1(D = 1)}{p(X)} \mid p(X)\right] \\
&= \mathrm{E}\left[g(Y)\frac{1(D = 1)}{p(X)} \mid D = 1, p(X)\right] P(D = 1|p(X)) \\
&\quad + \mathrm{E}\left[g(Y)\frac{1(D = 1)}{p(X)} \mid D = 0, p(X)\right] P(D = 0|p(X)) \\
&= \mathrm{E}[g(Y) \mid D = 1, p(X)]\frac{P(D = 1 \mid p(X))}{p(X)} \\
&= \mathrm{E}[g(Y) \mid D = 1, p(X)] \\
&= \mathrm{E}[g(Y(1)) \mid D = 1, p(X)]
\end{aligned}$$

where we use $P(D = 1 \mid p(X)) = p(X)$. We can similarly argue for the case of $d = 0$. Thus, the conditional distribution of $Y(1)$ does not depend on D, once we condition on $p(X)$, which verifies Theorem 5.4.1.

5.B Clever Covariate Regression

Here we show that if we care only about estimating the ATE, then it suffices to learn the BLP of the outcome Y using the single covariate

$$\phi(D, X) := H = \frac{1(D = 1)}{p(X)} - \frac{1(D = 0)}{1 - p(X)}.$$

We can then use this BLP model as a proxy for the CEF $\mathrm{E}[Y \mid D, p(X)]$. Specifically, we learn a decomposition $Y = \beta\phi(D, X) + \epsilon, \epsilon \perp \phi(D, X)$ by running OLS of Y on $\phi(D, X)$ and then use $\mathrm{E}[\beta(\phi(1, X) - \phi(0, X))]$ as the ATE. This approach, referred to in the literature as the "clever covariate" approach, was first proposed in [7] and further developed in [8].

Note that the random variable H satisfies

$$\mathrm{E}[f(D, X)H \mid X] = f(1, X) - f(0, X)$$

for any function $f(D, X)$.[9] Then, by Theorem 5.3.1 and orthogonality of ϵ in the BLP decomposition:

$$\begin{aligned} \mathrm{E}[Y(1) - Y(0)] &= \mathrm{E}[YH] = \mathrm{E}[\beta\phi(D, X)H] \\ &= \mathrm{E}\left[\beta(\phi(1, X) - \phi(0, X))\right]. \end{aligned}$$

Note that even though this approach allows us to identify the ATE, it does uncover the CATE $\mathrm{E}[Y(1) - Y(0) \mid X]$. The reason for the failure in learning the CATE is that the residual ϵ does not necessarily satisfy conditional orthogonality; i.e. we do not have $\mathrm{E}[(Y - \beta\phi(D, X))H \mid X] = 0$.

5.C Details of ATET

In observational studies, the ATET is identified under weaker conditions than the ATE because

$$\mathrm{E}[Y(1) \mid D = 1] = \mathrm{E}[Y \mid D = 1],$$

so we only need to identify $\mathrm{E}[Y(0) \mid D = 1]$. We can state the weaker version of the ignorability and overlap conditions as follows:

Assumption 5.C.1 (Ignorability and Overlap for Treated) *(a) Ignorability. Suppose that the treatment status D is independent of*

[9] Verify this as a reading exercise.

$Y(0)$ *conditional on a set of covariates* X*, that is*

$$D \perp\!\!\!\perp Y(0) \mid X.$$

(b) Weak Overlap. Suppose that the propensity score satisfies:

$$\mathrm{P}(p(X) < 1) = 1.$$

Theorem 5.C.1 (Identification of ATET) *Under Assumption 5.C.1,*

$$\delta_1 = \mathrm{E}[Y \mid D = 1] - \mathrm{E}[\mathrm{E}[Y \mid X, D = 0] \mid D = 1].$$

Theorem 5.C.1 follows because, by iterated expectations and ignorability,

$$\begin{aligned} \mathrm{E}[Y(0) \mid D = 1] &= \mathrm{E}[\mathrm{E}[Y(0) \mid D = 1, X] \mid D = 1] \\ &= \mathrm{E}[\mathrm{E}[Y(0) \mid D = 0, X] \mid D = 1] \\ &= \mathrm{E}[\mathrm{E}[Y \mid D = 0, X] \mid D = 1], \end{aligned}$$

where the outer expectation is well-defined because the support of X conditional on $D = 1$ is a subset of the support of X conditional on $D = 0$ by the overlap condition.

The Horvitz-Thompson method can be also used to recover averages of potential outcomes for the treated. Indeed,

$$\frac{\mathrm{E}[DY]}{\mathrm{E}[D]} = \frac{\mathrm{E}[DY(1)]}{\mathrm{E}[D]} = \mathrm{E}[Y(1) \mid D = 1]$$

and

$$\begin{aligned}
\frac{\mathrm{E}\left[\frac{(1-D)}{1-p(X)}p(X)Y\right]}{\mathrm{E}[D]} &= \frac{\mathrm{E}\left[\frac{p(X)}{1-p(X)}\mathrm{E}[(1-D)Y \mid X]\right]}{\mathrm{E}[D]}\\
&= \frac{\mathrm{E}\left[\frac{p(X)}{1-p(X)}\mathrm{E}[(1-D)Y(0) \mid X]\right]}{\mathrm{E}[D]}\\
&= \frac{\mathrm{E}\left[\frac{p(X)}{1-p(X)}\mathrm{E}[1-D|X]\mathrm{E}[Y(0) \mid X]\right]}{\mathrm{E}[D]}\\
&= \frac{\mathrm{E}[p(X)\mathrm{E}[Y(0) \mid X]]}{\mathrm{E}[D]}\\
&= \frac{\mathrm{E}[\mathrm{E}[D \mid X]\mathrm{E}[Y(0) \mid X]]}{\mathrm{E}[D]}\\
&= \frac{\mathrm{E}[\mathrm{E}[DY(0) \mid X]]}{\mathrm{E}[D]}\\
&= \frac{\mathrm{E}[DY(0)]}{\mathrm{E}[D]} = \mathrm{E}[Y(0) \mid D = 1]
\end{aligned}$$

where in the second to last step we used that $D \perp\!\!\!\perp Y(0) \mid X$, implies $\mathrm{E}[DY(0) \mid X] = \mathrm{E}[D|X]\mathrm{E}[Y(0) \mid X]$. Hence, we obtain the following result:

Theorem 5.C.2 (Propensity Score Reweighting for the Treated) *Under Assumption 5.C.1,*

$$\mathrm{E}[Y\bar{H}] = \delta_1, \quad \bar{H} = Hp(X)/\mathrm{E}[D].$$

Bibliography

[1] Merriam Webster Dictionary. *Compare apples and/to/with apples*. URL: https://www.merriam-webster.com/dictionary/compare%20apples%20and%2Fto%2Fwith%20apples (cited on page 129).

[2] Donald B. Rubin. 'Estimating causal effects of treatments in randomized and nonrandomized studies.' In: *Journal of Educational Psychology* 66.5 (1974), pp. 688–701 (cited on page 131).

[3] Daniel G. Horvitz and Donovan J. Thompson. 'A generalization of sampling without replacement from a finite universe'. In: *Journal of the American Statistical Association* 47.260 (1952), pp. 663–685 (cited on page 137).

[4] Atila Abdulkadiroğlu, Joshua D. Angrist, Yusuke Narita, and Parag A. Pathak. 'Research Design Meets Market Design: Using Centralized Assignment for Impact Evaluation'. In: *Econometrica* 85.5 (2017), pp. 1373–1432. DOI: 10.3982/ECTA13925 (cited on page 138).

[5] Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val. *Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in India*. Tech. rep. National Bureau of Economic Research, 2018 (cited on page 140).

[6] Paul R. Rosenbaum and Donald B. Rubin. 'The Central Role of the Propensity Score in Observational Studies for Causal Effects'. In: *Biometrika* 70.1 (1983), pp. 41–55 (cited on page 140).

[7] Daniel O. Scharfstein, Andrea Rotnitzky, and James M. Robins. 'Adjusting for Nonignorable Drop-Out Using Semiparametric Nonresponse Models'. In: *Journal of the American Statistical Association* 94.448 (1999), pp. 1096–1120. (Visited on 01/25/2023) (cited on pages 142, 145).

[8] Heejung Bang and James M. Robins. 'Doubly Robust Estimation in Missing Data and Causal Inference Models'. In: *Biometrics* 61.4 (2005), pp. 962–973. DOI: https://doi.org/10.1111/j.1541-0420.2005.00377.x (cited on pages 142, 145).